2015-2017 National Survey of Family Growth Male Questionnaire Datasheet

Sophia Atik
EM 0212- Applied Data Science
March 10th, 2019

User Guide: https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf
Questionnaire: https://www.cdc.gov/nchs/data/nsfg/NSFG_2015-2017_MaleCAPIlite_forPUF.pdf
Datasets: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

This dataset was created for the National Survey of Family Growth to understand the male perspective of contraceptives uses in a house hold, contraceptive uses between partners, sex communication, past and current sexual partners, past and current cohabiting partners, number of children (adopted or biological), what kind of father the interviewee is and his desire to have future children, demographics, and other information about his sex life.

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

This dataset should NOT be used to try to figure out who the people interviewed were. It could be used to figure out what certain groups of men use certain contraceptives, why some men get vasectomies, and what outside information influences their decisions.

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Current Contraceptive Use and Variation by Selected Characteristics Among Women Aged 15–49: United States, 2015–2017 https://www.cdc.gov/nchs/products/databriefs/db327.htm
(However I am using the male questionnaire specifically for my dataset from the years 2015-2017)

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics

Any other comments?
N/A

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)
People

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?
N/A

How many instances of each type are there?
One (interviewee = one source = one instance (I think))

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?
Unprocessed raw data.

ORGANIZATION OF THE 2015-2017 NSFG PUBLIC-USE DATA FILES

The public-use data for the 2015-2017 NSFG are provided as three separate ASCII files.

FILE CHARACTERISTICS	**Number of Records (observations)**	**Record Length (number of columns)**	**Number of Variables**
Female respondent file File = 2015_2017_FemRespData.dat (one record per female respondent)	5,554	4,522	3,024
Female pregnancy (interval) file File = 2015_2017_FemPregData.dat (one record per pregnancy reported by female respondents)	9,553	375	248
Male respondent file File = 2015_2017_MaleData.dat (one record per male respondent)	4,540	4,170	2,945

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with *any* of the data?

Everything is included but must be accessed by an external link.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

There is a Data User's Agreement that says I must comply to with the following legal requirements:

- "To use these data for statistical reporting and analysis only;
- To make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (301-458-4500); and
- To not link these data with individually identifiable data from any other data set."

https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf

What experiments were initially run on this dataset?

Have a summary of those results and, if available, provide the link to a paper with more information here.

I have not found any papers using specifically the male questionnaire data. There is a paper using the female questionnaire data to discover what percentage of women use what contraceptives. https://www.cdc.gov/nchs/products/databriefs/db327.htm

Any other comments?

N/A

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software pro- gram, software interface/API; how were these con- structs/measures/methods validated?)

The data was collected from 60 minute in-person interviews. Some answers were also documented via a computer at the end of the in-person interview by the interviewee.
Interviews were conducted under Research Ethics Review Board guidelines
Unknown how they were validated.

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Interviewees were incentivized with $40. I believe it was someone's job to collect this data from the National Center for Health Statistics, which is an agency within in the U.S. Department of Health and Human Services' Centers for Disease Control and Prevention (DHHS/CDC).

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

60 minute interviews were collected from male interviewees from September 2015 to September 2017.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Observable in-person interviews and survey responses reported by subjects

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

I think it is a sample of a larger set of instances. It says that the data over-samples African American men, Hispanic men, and teenage boys.

If the dataset is a sample, then what is the population?

What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of in- stances)? How does this affect possible uses?

There were 4,540 men interviewed between the ages 15-49. The total population would be all men 15-49 of all races in the United States household population.

Is there information missing from the dataset and why?

(this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

> "NCHS does all it can to assure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the data files. In addition, some records have had one or more responses slightly modified through statistical perturbation. These modifications are intended to prevent definitive identification of individual respondents. They do not affect univariate point estimates and have a minimal effect on estimates of variance and tests of statistical significance."

https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf

Are there any known errors, sources of noise, or redundancies in the data?

I cannot tell yet.

Any other comments?
N/A

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)
Unknown, but they do seem to have variance estimation examples provided.
https://www.cdc.gov/nchs/nsfg/nsfg_2015_2017_puf.htm#informed

Was the "raw" data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)
Unknown

Is the preprocessing software available?
Unknown

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?
Unknown

Any other comments?
When I open the dataset, all of the variables are listed in the same column. However, the User Guide states:

> "The Male Respondent file, often just referred to as the male file, contains one record for each of the 4,540 men interviewed; the male respondent is the unit of analysis. Program statements are provided on the NSFG webpage to read the ASCII data into SAS, SPSS, and Stata, and include variable and value labels. Formats are provided within the program statements, but are for user convenience or ease of display only. They do not always reflect the actual values in the dataset and sometimes condense responses into categories. Data users should make their own decisions about whether to use the formats provided in the program statements. In addition, SAS and Stata syntax guidelines are provided in Appendix 2 for combining data file releases."
> https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf

I am confused and am concerned that I am not accessing or opening the file correctly. I searched the User Guide to understand exactly what kind of preprocessing was done. I could not find an answer

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

Website ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

December 2018

What license (if any) is it distributed under? Are there any copyrights on the data?

No. They just require proper citation of the data:

> National Center for Health Statistics (NCHS). (2018). 2015-2017 National Survey of Family Growth Public-Use Data and Documentation. Hyattsville, MD: CDC National Center for Health Statistics. Retrieved from **http://www.cdc.gov/nchs/nsfg/nsfg_2015_2017_puf.htm.**

Are there any fees or access/export restrictions?

No

Any other comments?

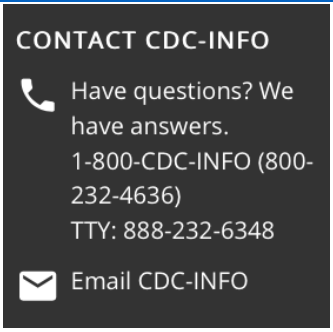
N/A

Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

The CDC is supporting the dataset. https://www.cdc.gov/nchs/nsfg/nsfg_2015_2017_puf.htm



Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?

It seems like the CDC updates this information every year. They have data collected on this same topic from 2006 to 2017. They even go back to 1973, but the years following were not updated each year until 2006.

If the dataset becomes obsolete how will this be communicated?
Unknown, but most likely updated on the CDC website.

Is there a repository to link to any/all papers/systems that use this dataset?
https://www.cdc.gov/nchs/nsfg/nsfg_products.htm

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?
Unknown, the above website does not say.

Any other comments?
N/A

Legal and Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)
Yes. This dataset collects personal answers from men from the age of 15-49. They were informed about the data collection and were incentivized with $40. The interviews were conducted under protocols and informed consent procedures that were approved by the National Center for Health Statistics Research Ethics Review Board (protocol #2015-12).

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)
N/A

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)
Yes, procedures were approved by the National Center for Health Statistics Research Ethics Review Board (protocol #2015-12).

If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?
Yes, they were told what the intended use was, consented, and received $40. They were informed with a brochure and contact information for if they had additional questions (https://www.cdc.gov/nchs/data/nsfg/nsfg_question_and_answer_brochure.pdf).

Unknown what is in place for revoking consent in the future or what norms exist.

If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

The NCHS made sure that someone's identity could not be determined by the information provided in the dataset. Determining the identity of the interviewee is prohibited by law. Anyone who broke an interviewee's confidentiality could be "fined up to $250,000, lose their job, and/or be sent to prison."
https://www.cdc.gov/nchs/data/nsfg/nsfg_question_and_answer_brochure.pdf

If it relates to people, does it unfairly advantage or dis- advantage a particular social group? In what ways? How was this mitigated?

Unknown

If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

"NCHS does all it can to assure that the identity of data subjects cannot be disclosed."
- **https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf**

Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

Unknown, but I assume yes

Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

While this dataset does contain sensitive and personal information, the NCHS took steps to "prevent definitive identification of individual respondents."
https://www.cdc.gov/nchs/data/nsfg/NSFG_2015_2017_UserGuide_MainText.pdf

Does the dataset contain information that might be considered inappropriate or offensive?

No. The questions can be personal due to the fact they are about people's intimate lives. However, if an interviewee preferred not to answer, they could "refuse." The most personal questions were asked at the end of the interview and the interviewee recorded the information via a survey on a computer.

Any other comments?

N/A